FROM CELLULAR RESPONSES TO PHARMACOLOGICAL DOMAINS: MULTIMODAL ZERO-SHOT DRUG REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal drug discovery enables drug representation learning beyond chemical structure by incorporating cellular responses such as gene expression and cell morphology. However, direct fusion and instance-level contrastive alignment may mix mechanism-related signals with modality-specific noise and incorrectly separate structurally dissimilar but biologically related compounds. This limitation can obscure transferable mechanism patterns required for predicting the properties of unseen compounds. We introduce PMRD, a pharmacological response domain-guided framework for multimodal zero-shot drug property prediction. PMRD separates mechanism-consistent factors from modality-specific information and constructs a consensus response domain across three modalities. Mechanism candidate augmentation identifies locally stable factors, while retrieval-geometry attribution dynamically reweights the alignment and augmentation objectives according to whether their updates preserve inter-drug discriminability. This feedback suppresses training signals that conflict with mechanism-discriminative retrieval. PMRD further combines complementary representations through reliability-aware multiview retrieval. Experiments on public datasets show improved zero-shot property prediction and more biologically coherent drug neighborhoods. Hard-negative analysis further indicates fewer conflicts between structurally dissimilar but response-related compounds. These results support PMRD as an effective framework for mechanism-aware multimodal drug representation learning.

1 INTRODUCTION

In recent years, deep learning has significantly advanced candidate compound screening, drug property prediction, and mechanism-of-action analysis in computer-aided drug discovery Rong et al. (2020). Drug property prediction aims to infer the biological effects of compounds from limited observations. However, relying solely on molecular structures is insufficient to fully characterize the complex drug action process Subramanian et al. (2017); Bray et al. (2017).

To alleviate this information bottleneck, existing multimodal learning approaches have begun to integrate molecular structures with biological response data, such as gene expression profiles and cellular morphology. As illustrated in Fig. 1, these complementary observations describe both the chemical identity of a compound and its induced cellular responses, providing a broader basis for mechanism-aware drug representation learning Wang et al. (2024).

Although multimodal information provides complementary perspectives for drug representation learning, its effective utilization remains limited by label scarcity and spurious correlations Wu et al. (2018); Hu et al. (2020); Wang et al. (2024). High-quality drug–phenotype labels usually depend on expensive wet-lab experiments and are particularly scarce for novel compounds, new targets, and unseen tasks, making it difficult for conventional supervised models to generalize to unseen scenarios Corsello et al. (2020). Meanwhile, molecular structures, gene expression profiles, and cellular morphology simultaneously contain pharmacologically relevant information, modality-specific information, and experimental noise Subramanian et al. (2017); Bray et al. (2017); McQuin et al. (2018); Wang et al. (2024). Directly fusing these observations may cause the model to mistake

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

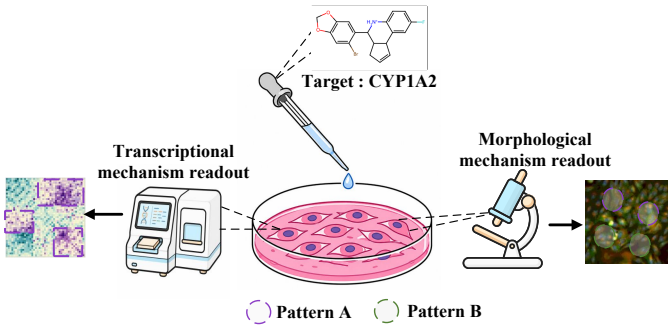


Figure 1: Conceptual overview of mechanism-aware drug profiling by integrating molecular structure with drug-induced transcriptional and morphological responses.

dataset biases or incidental co-occurrences for valid pharmacological signals, thereby weakening its transferability in zero-shot scenarios Du et al. (2023); Wang et al. (2024).

Therefore, the key challenge in multimodal zero-shot drug representation learning is not merely to enhance statistical consistency across modalities, but to identify stable response factors that reflect drug action mechanisms from complex multi-source observations. This motivation is supported by prior findings that contrastive learning may rely on shortcut features Robinson et al. (2021), high-throughput drug screens often contain batch-induced non-biological associations Wang et al. (2024), and multimodal models can suffer from modality dominance, where cleaner modalities suppress weaker but biologically informative views Du et al. (2023). Moreover, weak molecule–phenotype alignment may make conventional negative sampling unreliable in phenotypic contrastive learning Rao et al. (2025). Therefore, the model needs to determine which information consistently supports drug property discrimination across different modality views, sample neighborhoods, and task conditions, and which signals mainly arise from noise, bias, or local co-occurrence under specific data distributions. Based on this consideration, we focus on the following question:

> *How can stable drug response factors be identified from multi-source observations for reliable zero-shot property prediction?*

To address this issue, we propose a multi-view zero-shot drug property prediction framework guided by a Pharmacological Mechanism Representation Domain (PMRD). Since each modality contains both shared drug responses and source-specific variations, Cross-Modal Response Modeling (CRM) decomposes its representation into a Mechanism-Consistent Representation (MCR) and a Modality-Specific Representation (MSR). MCR captures transferable pharmacological knowledge that remains consistent across heterogeneous modalities, whereas MSR preserves complementary modality-dependent characteristics. In this way, we can prevent modality-specific variations from interfering with mechanism-level alignment, thereby improving cross-modal knowledge transfer to unseen drug properties.

However, the above representations remain sensitive to variations, and to make it representation more robust and stable, we further propose a Stability-Guided Optimization (SGO) module. Specifically, SGO consists of the Candidate Augmentation Mechanism (CAM) and the Data Attribution Mechanism (DAM) modules. CAM attempts to regularize the MCRs using locally perturbed candidates that preserve semantic and cross-modal consistency. DAM further assesses whether the updates preserve inter-drug discriminability and adaptively adjusts their weights accordingly.

Finally, since no labeled data are available in the zero-shot setting, our framework cannot train a conventional prediction head. We therefore introduce a Reliability-Aware Retrieval (RAR) process that performs property inference at test time by retrieving reliable reference samples from the learned representation space. Specifically, RAR retrieves the top-K most similar drugs from the training set in each complementary representation view and aggregating their similarity-weighted predictions according to property-specific view reliability. This allows unseen drugs to exploit reliable property evidence without relying on a single fixed representation. Our main contributions are summarized as follows:

- We propose a PMRD-guided framework for multimodal zero-shot drug property prediction, shifting the focus from fixed multimodal fusion to learning and utilizing stable drug-response evidence from heterogeneous observations.
- We develop CRM and SGO to construct cross-modally supported drug representations and refine them toward local stability and inter-drug discriminability, thereby reducing the influence of modality-specific bias and unreliable training signals.
- We introduce RAR to adaptively aggregate multiple retrieval views according to their property-specific reliability, enabling the model to exploit informative cross-view evidence without relying on a single representation.

Code availability. The code will be released upon publication.

2 Related Work

2.1 Molecular Representation Learning

Molecular representation learning is a core problem in drug property prediction. Early methods use hand-crafted descriptors and fingerprints to encode local chemical substructures Rogers & Hahn (2010). Deep models further represent molecules as atom–bond graphs and learn structural features through message passing or attention mechanisms Gilmer et al. (2017); Yang et al. (2019); Xiong et al. (2020). Standard benchmarks such as MoleculeNet highlight the importance of scaffold split, label scarcity, and out-of-distribution generalization Wu et al. (2018).

To reduce reliance on task-specific labels, self-supervised molecular pre-training has been widely explored. Context prediction, node masking, edge prediction, and graph contrastive learning encourage transferable molecular representations before downstream fine-tuning Hu et al. (2020); You et al. (2020); Wang et al. (2022). Representative methods include GROVER, GraphLoG, and JOAO Rong et al. (2020); Xu et al. (2021); You et al. (2021). Other studies exploit 3D molecular geometry to capture spatial chemical information Zhou et al. (2023); Stärk et al. (2022). However, these methods mainly rely on molecular structures and miss drug response factors observable only from phenotypic readouts.

2.2 Multimodal Phenotypic Drug Representation Learning

Phenotypic screening provides complementary biological evidence for modeling drug responses. LINCS L1000 measures transcriptional perturbations, while Cell Painting captures morphology changes induced by compounds Subramanian et al. (2017); Bray et al. (2017); McQuin et al. (2018). Large-scale resources such as JUMP Cell Painting further support phenotype-aware drug discovery Chandrasekaran et al. (2023). These readouts have been used for mechanism-of-action analysis, drug repurposing, and phenotype-based retrieval Corsello et al. (2020); Jang et al. (2021).

Recent multimodal methods align molecular structures with phenotypic observations. Inspired by CLIP-style contrastive alignment Radford et al. (2021a), CLOOME aligns chemical structures with cellular images Sanchez-Fernandez et al. (2023b), MIGA performs graph–image contrastive learning Zheng et al. (2024b), InfoCORE reduces batch-effect bias via conditional mutual information maximization Wang et al. (2024), and MINER calibrates negative sampling for incomplete multimodal observations Rao et al. (2025). These methods validate the usefulness of phenotypic data, but mainly focus on cross-modal consistency or supervised downstream prediction, leaving stable mechanism-related factor selection under zero-shot settings less explored.

2.3 Zero-shot and Robust Multimodal Prediction

Zero-shot drug property prediction aims to infer the properties of unseen compounds without task-specific supervised fine-tuning by exploiting transferable representations learned from chemical structures and perturbation-derived phenotypes Sanchez-Fernandez et al. (2023a); Liu et al. (2025). In multimodal settings, this problem is complicated by the coexistence of pharmacologically relevant responses, modality-specific variations, batch effects, and experimental noise Zheng et al. (2024a); Wang et al. (2024). Optimization may also exhibit modality dominance, where stronger or cleaner

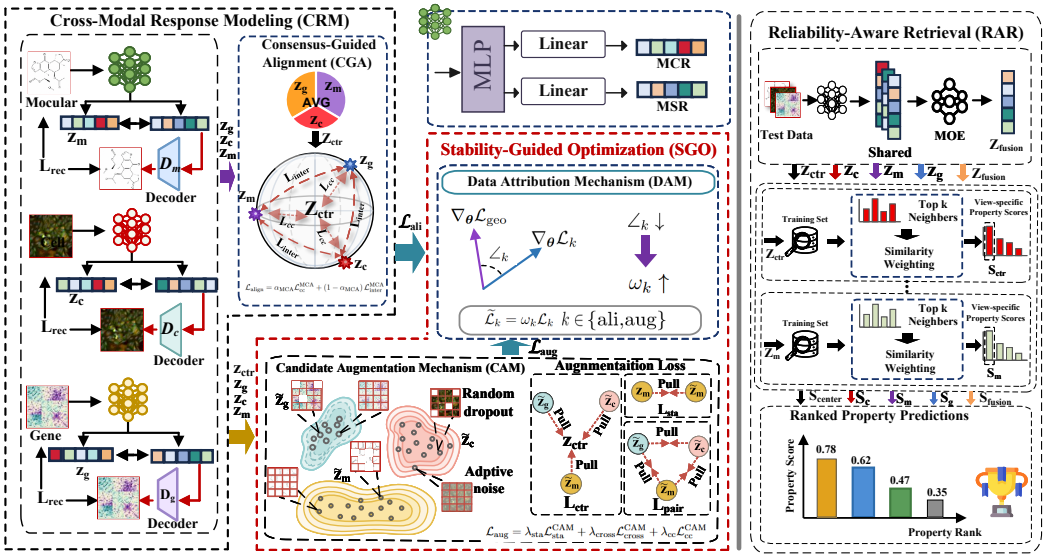


Figure 2: Overview of the proposed PMRD framework. PMRD disentangles multimodal drug observations into mechanism-consistent and modality-specific representations, constructs a cross-modal consensus mechanism domain, and uses attribution-guided loss reweighting to retain effective training signals. The learned representations are subsequently integrated through multiview retrieval for zero-shot drug property prediction.

modalities suppress weaker but biologically informative views Du et al. (2023). The central challenge is therefore to preserve transferable and complementary cross-modal evidence while limiting interference from source-specific variation, technical confounders, and dominant modalities.

3 Proposed Method

3.1 Overview

To identify stable mechanistic evidence from heterogeneous drug observations, we propose a PMRD-guided framework for multimodal zero-shot drug property prediction, as illustrated in Fig. 2. Given molecular structure, gene expression, and cellular morphology, the framework proceeds through CRM, SGO, and RAR. CRM directs shared drug responses and source-specific variations into the MCR and MSR, respectively, preventing modality bias from disrupting cross-modal knowledge transfer. CGA then constructs a drug-specific PMRD consensus center from the three MCRs and aligns them with this reference while preserving inter-drug distinctions. Because the shared responses identified by CRM may depend on incidental features, SGO uses CAM to retain factors whose semantics and cross-modal relations persist under local perturbations. DAM further adjusts the weights of CGA and CAM according to whether their updates preserve inter-drug discriminability. During zero-shot prediction, task-specific fine-tuning is unavailable and view reliability varies across properties. RAR adaptively aggregates multi-view predictions based on the reliability of different attributes from each view, avoiding dependence on a single fixed representation.

3.2 Cross-Modal Response Modeling

Drug observations mix cross-modal response patterns with modality-specific variations, so directly aligning them may introduce source-dependent bias. CRM separates these factors and consolidates response information supported across modalities.

For drug $i \in \mathcal{B}$ in modality m, where $\mathcal{B}$ is the current mini-batch and $\mathbf{x}_i^m$ is its modality input, two projection heads produce MCR $\mathbf{z}_i^m$ and MSR $\mathbf{u}_i^m$, with $m \in \mathcal{M} = \{\mathrm{mol}, \mathrm{gene}, \mathrm{cell}\}$. The MCR participates in CGA to retain transferable response information, whereas the MSR reconstructs its

own modality to preserve source-specific variation. Their redundancy is reduced by

$$\mathcal{L}_{\text{sep}} = \text{Avg}_{i,m}\left[\text{sim}(\mathbf{z}_i^m, \mathbf{u}_i^m)^2\right]. \tag{1}$$

Here, $\text{Avg}_{i,m}$ averages over $i \in \mathcal{B}$ and $m \in \mathcal{M}$, and $\text{sim}(\cdot,\cdot)$ denotes cosine similarity. Minimizing Eq. equation 1 reduces information overlap between the two branches, while reconstruction prevents trivial separation through information loss.

Because a single MCR may retain correlations supported by only one modality, CGA averages and renormalizes the three MCRs of each drug to construct a drug-specific PMRD center. This operation strengthens shared response directions while suppressing unsupported variations. CGA combines consensus-center and inter-modal alignment:

$$\mathcal{L}_{\text{ali}} = \alpha_{\text{CGA}}\mathcal{L}_{\text{cc}}^{\text{CGA}} + (1-\alpha_{\text{CGA}})\,\mathcal{L}_{\text{inter}}^{\text{CGA}}. \tag{2}$$

Here, $\mathcal{L}_{\text{ali}}$ is the overall CGA loss, $\mathcal{L}_{\text{cc}}^{\text{CGA}}$ aligns each MCR with the PMRD center of the same drug, and $\mathcal{L}_{\text{inter}}^{\text{CGA}}$ directly aligns its three MCRs to prevent center averaging from masking cross-modal disagreement. Other drugs in the mini-batch serve as negatives to preserve inter-drug discriminability. The coefficient $\alpha_{\text{CGA}} \in [0,1]$ balances the two terms and is set to 0.6.

3.3 Stability-Guided Optimization

The cross-modal responses identified by CGA may rely on incidental features and disappear under mild perturbations. Moreover, CGA and CAM may contribute unequally across mini-batches, while excessive alignment or augmentation can weaken inter-drug discriminability. SGO therefore uses CAM to retain locally stable factors and DAM to select optimization signals that preserve drug-level distinctions.

CAM perturbs each MCR through random feature dropout and scale-adaptive Gaussian noise. It retains candidates that preserve semantics, cross-modal relations, PMRD-center consistency, and local distribution. Samples without valid candidates contribute no augmentation loss. The retained candidates regularize the original MCRs through

$$\mathcal{L}_{\text{aug}} = \lambda_{\text{sta}}\mathcal{L}_{\text{sta}}^{\text{CAM}} + \lambda_{\text{cross}}\mathcal{L}_{\text{cross}}^{\text{CAM}} + \lambda_{\text{cc}}\mathcal{L}_{\text{cc}}^{\text{CAM}}, \tag{3}$$

where $\mathcal{L}_{\text{sta}}^{\text{CAM}}$ preserves candidate semantics and local structure, $\mathcal{L}_{\text{cross}}^{\text{CAM}}$ maintains cross-modal consistency, and $\mathcal{L}_{\text{cc}}^{\text{CAM}}$ aligns candidates with the corresponding PMRD center. The coefficients λ_{sta}, λ_{cross}, and λ_{cc} are fixed nonnegative weights. The candidates regularize local MCR neighborhoods but neither replace the original MCRs nor participate in retrieval.

Because local stability alone does not ensure inter-drug separation, DAM defines an attribution reference using the multi-view retrieval geometry:

$$\mathcal{L}_{\text{geo}} = \text{Avg}_{v\in\mathcal{V},\, i\neq j}\left[\text{sim}\big(\mathbf{h}_i^v, \mathbf{h}_j^v\big)^2\right], \tag{4}$$

where $\mathcal{V} = \{\text{mol}, \text{gene}, \text{cell}, \text{ctr}, \text{fus}\}$ contains the three MCR views, the PMRD-center view, and the fusion view. The vector $\mathbf{h}_i^v$ represents drug $i \in \mathcal{B}$ in view $v \in \mathcal{V}$, while $i \neq j$ excludes self-similarity. The averaging covers all views and distinct drug pairs in the current mini-batch. Minimizing Eq. equation 4 suppresses excessive similarity between different drugs, providing a reference direction for evaluating CGA and CAM updates.

DAM converts their directional support into dynamic weights:

$$\widetilde{\mathcal{L}}_k = \omega_k \mathcal{L}_k. \tag{5}$$

Here, $k \in \{\text{ali}, \text{aug}\}$ indexes the CGA and CAM objectives, ω_k is the dynamic weight assigned by DAM, and $\widetilde{\mathcal{L}}_k$ is the reweighted objective. Updates supporting inter-drug discriminability receive larger weights, whereas conflicting signals are suppressed. $\mathcal{L}_{\text{geo}}$ is used only as an attribution reference and is not added to the training loss; reconstruction retains a fixed weight. SGO thus favors response factors that are both locally stable and discriminative across drugs.

3.4 Reliability-Aware Retrieval

Because property labels are unavailable for unseen query drugs, the zero-shot protocol precludes task-specific fine-tuning. RAR instead retrieves similar labeled drugs from the training set. Since a view may be reliable for one property but less informative for another, it independently evaluates the three MCRs, the PMRD center, and the MoE fusion representation.

In each view, RAR retrieves the top-K training drugs by cosine similarity, where K is the number of neighbors, and computes a similarity-weighted prediction from their valid labels. Missing labels are excluded; if no retrieved drug is labeled for a property, its positive-label proportion in the training set is used as the fallback score.

RAR aggregates the view-specific predictions according to property-specific reliability:

$$\hat{y}_t(q) = \sum_{v \in \mathcal{V}} \rho_{v,t} \hat{y}_{v,t}(q), \tag{6}$$

where q denotes a query drug, t indexes a property, and $v \in \mathcal{V}$ indexes a retrieval view. $\hat{y}_{v,t}(q)$ is the prediction from view v, and $\rho_{v,t}$ is its nonnegative reliability weight normalized over $\mathcal{V}$. Reliability is estimated from prediction confidence, cross-view agreement, neighborhood overlap, label coverage, and the attribution prior. The learned representations determine which training drugs are retrieved, while RAR determines which views are reliable for each property.

4 Experiments

4.1 Experimental Setup

4.1.1 Datasets

We conduct experiments on two multimodal drug property datasets, ChEMBL2K Gaulton et al. (2012) and Broad6K Moshkov et al. (2023). ChEMBL2K contains 2,355 drug samples and 41 binary property prediction tasks, while Broad6K contains 6,567 drug samples and 32 binary property prediction tasks. Each sample includes three types of observations: molecular structure, gene expression, and cellular morphology. To evaluate predictive performance on unseen drugs, we adopt a scaffold split and partition each dataset into training, validation, and test sets with a ratio of $0.60/0.15/0.25$. ChEMBL2K is divided into 1,413 training, 353 validation, and 589 test samples, respectively. Broad6K is divided into 3,940 training, 985 validation, and 1,642 test samples, respectively.

4.1.2 Evaluation Metrics

We use the area under the receiver operating characteristic curve (AUROC) as the primary evaluation metric. In addition to mean AUROC across tasks, we report the proportions of tasks achieving AUROC values above 0.80, 0.85, and, where applicable, 0.90. Unless otherwise specified, results are reported as the mean and standard deviation over five independent runs.

4.1.3 Baseline

We compare our method with four groups of baseline methods. The supervised unimodal group contains models based on gene expression, cellular morphology, and Morgan fingerprints. The GNN-pretraining group includes AttrMask, EdgePred Hu et al. (2020), GROVER Rong et al. (2020), GraphLoG Xu et al. (2021), and JOAO You et al. (2021). The multimodal-alignment group includes CLOOME Sanchez-Fernandez et al. (2023a), MIGA Zheng et al. (2024a), and InfoCORE Wang et al. (2024). The zero-shot alignment group contains CLIP Radford et al. (2021b), CCL Ma et al. (2022), and PMRD (Ours).

4.1.4 Implementation Details

The molecular structure encoder is implemented as a three-layer Graph Isomorphism Network (GIN), whereas the gene-expression and cellular-morphology encoders are implemented as

Table 1: Molecular property prediction on ChEMBL2K and Broad6K. We report mean AUROC and the proportion of valid tasks exceeding each AUROC threshold. Best and second-best results are shown in **bold** and underline, respectively; † denotes zero-shot methods.

Category	Method	ChEMBL2K				Broad6K		
		Avg.	$> 80\%$	$> 85\%$	$> 90\%$	Avg.	$> 80\%$	$> 85\%$
Supervised	Gene Expression	56.1±1.1	5.1±1.4	3.4±1.3	3.4±1.3	56.9±1.4	1.9±1.7	1.9±1.7
	Cell Morphology	64.3±2.4	15.6±6.6	8.3±3.7	4.9±3.9	55.3±0.1	0.0±0.0	0.0±0.0
	Morgan FP	76.8±2.2	48.8±3.9	34.6±6.3	21.9±5.7	63.3±0.3	6.3±0.0	**4.4±1.7**
GNN Pretraining	AttrMask	73.3±1.4	46.3±3.4	30.3±2.4	14.6±1.7	59.8±0.2	3.1±0.0	3.1±0.0
	EdgePred	75.6±0.5	54.2±4.0	34.6±7.2	12.2±2.4	59.9±0.2	3.1±0.0	3.1±0.0
	GROVER	73.3±1.4	38.5±2.0	22.4±3.6	14.6±2.4	66.2±0.1	**15.6±0.0**	3.8±1.4
	GraphLoG	73.5±0.7	41.9±2.0	29.3±3.4	15.6±2.8	62.9±0.4	4.4±1.7	0.0±0.0
	JOAO	75.1±1.0	47.8±5.1	33.7±2.0	19.0±3.2	67.3±0.4	12.5±0.0	3.8±1.4
Multimodal Alignment	CLOOME	66.7±1.8	26.8±4.6	16.1±3.7	10.7±5.1	61.7±0.4	3.1±0.0	3.1±0.0
	MIGA	75.6±1.6	**55.1±1.3**	35.1±4.4	14.6±1.7	62.9±0.4	4.4±1.7	0.0±0.0
Zero-shot Alignment	CLIP†	71.1±0.8	37.0±2.8	20.9±4.5	15.1±3.9	67.4±0.3	9.3±1.9	3.1±0.0
	CCL†	71.5±0.8	33.1±5.0	22.9±4.7	15.6±2.4	67.2±0.5	10.0±2.3	3.1±0.0
	Ours†	**77.8±0.3**	54.6±3.7	**37.1±3.2**	**25.4±3.7**	**67.9±0.5**	11.2±3.6	**4.4±1.7**

five-layer deep neural networks. The hidden dimension is 64. We train the model for 240 epochs with a batch size of 1024 and a learning rate of 0.005.

During zero-shot retrieval, we use the top-20 nearest neighbors and set the retrieval temperature to 0.1. For PMRD alignment, the weights of reference alignment and inter-modality alignment are 0.6 and 0.4, respectively. The stable candidate augmentation module generates eight local candidates, and its augmentation weight is 0.3. During inference, we retain five complementary views: the fusion view, molecular MCR view, gene MCR view, cellular MCR view, and PMRD view. Their predictions are aggregated using task-level weights based on view reliability.

4.2 Molecular Property Prediction

We evaluate PMRD under both strict zero-shot and label-informed settings. In the strict zero-shot setting, target-property labels are excluded from representation learning, and predictions are obtained solely through reliability-aware retrieval over the five learned views. Table 1 compares PMRD with supervised unimodal, GNN-pretraining, multimodal-alignment, and zero-shot baselines on ChEMBL2K and Broad6K.

On ChEMBL2K, PMRD achieves the highest average AUROC of 77.8%, exceeding the strongest supervised unimodal baseline, Morgan FP, by 1.0 percentage point. It also obtains the best proportions of tasks above the 0.85 and 0.90 thresholds, reaching 37.1% and 25.4%, respectively. Although MIGA is slightly higher at the 0.80 threshold, PMRD performs better at the stricter thresholds, indicating that its gains extend beyond marginal improvements and benefit a larger number of high-performing tasks.

On Broad6K, PMRD again achieves the highest average AUROC of 67.9%, outperforming CLIP and JOAO by 0.5 and 0.6 percentage points, respectively. It also ties for the best proportion of tasks above 0.85. While PMRD does not lead at the 0.80 threshold, its superior average performance suggests a more balanced improvement across properties rather than gains concentrated in a small subset of tasks. These results support the central motivation of PMRD: separating modality-specific variation, retaining locally stable response factors, and selecting reliable retrieval views can improve the transferability of multimodal drug representations to unseen property tasks without task-specific supervision.

Extension with Property-Label Supervision. We further evaluate a label-informed variant by adding a masked prediction loss over observed training labels while retaining the original PMRD objectives.

Table 2: Label-informed AUROC comparison on ChEMBL2K. Best results are shown in **bold**.

Method	Avg.	$> 80\%$	$> 85\%$	$> 90\%$
InfoCORE (CP)	73.8±2.0	37.6±9.2	26.3±4.7	10.7±4.1
InfoCORE (GE)	79.3±0.9	62.4±2.8	46.3±3.0	30.3±2.2
MINER	81.2±1.0	63.9±3.2	50.2±4.2	**37.5±4.5**
Ours variant	**81.6±1.0**	**67.8±2.8**	**54.6±3.7**	36.6±4.4

Table 3: Same-target rate (%) among Top-K candidates with Morgan Tanimoto similarity below 0.1. Lower is better.

Method	Top-10 ↓	Top-20 ↓	Top-30 ↓
CLIP	23.60%	23.63%	23.00%
CCL	22.73%	22.34%	22.11%
InfoCORE (CP)	22.15%	20.77%	20.82%
MINER	27.75%	29.30%	28.75%
Ours	**10.69%**	**10.38%**	**10.83%**

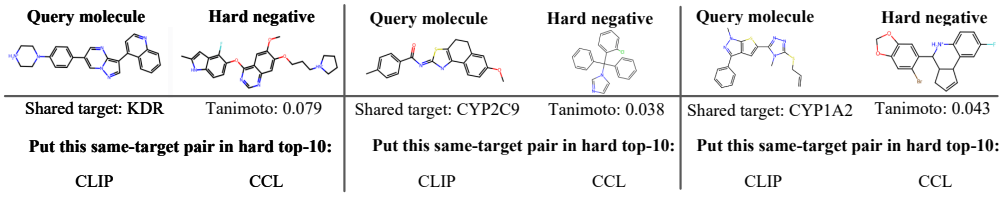


Figure 3: Representative low-structure hard-negative retrieval cases.

As shown in Table 2, label supervision raises the average AUROC from 77.8% to 81.6% and yields the best results at the 0.80 and 0.85 thresholds. This indicates that PMRD representations remain effective when task labels are available, while retaining their zero-shot utility when labels are absent.

4.3 Low-Structure Hard-Negative Analysis

Instance-level contrastive alignment may treat structurally dissimilar but biologically related compounds as negatives. To quantify this effect, we retain compound pairs with Morgan Tanimoto similarity below 0.1 and measure the same-target rate among the top-K non-self candidates, where a lower value indicates fewer target-overlapping negatives.

As shown in Table 3, PMRD achieves 10.38–10.83% across retrieval depths, substantially below the 20.77–29.30% range of the baselines. This suggests that PMRD is less likely to separate structurally dissimilar compounds that share annotated targets.

Fig. 3 further shows that several baseline-retrieved compounds share targets such as KDR, CYP2C9, and CYP1A2 despite low structural similarity. These cases illustrate potential conflicts in structure-driven negative selection. Since target annotations are incomplete, we interpret this analysis as evidence of reduced target conflicts rather than complete mechanism equivalence.

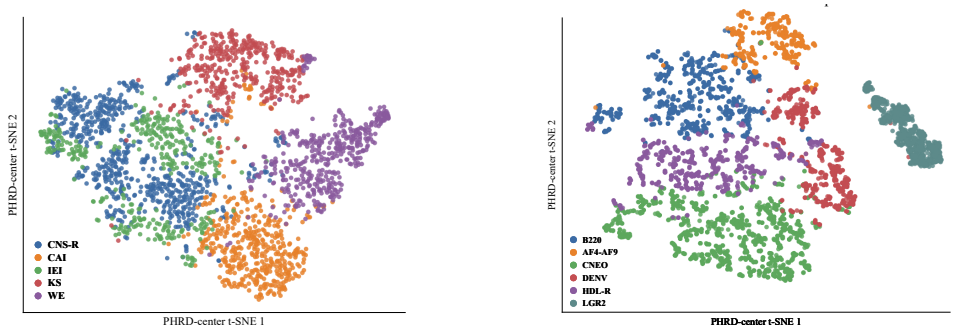
4.4 Ablation Study

We evaluate the contributions of MCR–MSR disentanglement, CGA, CAM, DAM, and RAR on ChEMBL2K.

Table 4 shows that removing MCR–MSR disentanglement or CGA reduces the average AUROC by 3.6 and 3.4 percentage points, respectively, supporting the roles of modality separation and consensus alignment. Removing CAM lowers the AUROC to 74.9%, indicating that local stability also contributes to mechanism representation learning. Without DAM, the average AUROC decreases to

Table 4: Ablation results on ChEMBL2K. The last two columns report the percentage of tasks exceeding the indicated AUROC thresholds.

Setting	Avg.	$> 80\%$	$> 85\%$
w/o MCR–MSR	74.2±1.8	43.4±7.1	33.2±4.0
w/o CGA	74.4±1.1	43.9±3.4	33.2±4.0
w/o CAM	74.9±1.2	44.4±3.6	34.6±4.7
w/o DAM	74.6±2.1	45.9±6.1	29.3±4.9
w/o RAR	67.8±1.9	21.0±1.2	15.1±1.8
Ours	**77.8±0.3**	**54.6±3.7**	**37.1±3.2**



(a) ChEMBL2K (b) Broad6K

Figure 4: t-SNE visualization of PMRD-center representations on ChEMBL2K and Broad6K. Colors denote enriched biological or assay themes.

74.6%, and the proportion of tasks above 85% drops by 7.8 percentage points. RAR has the largest observed effect: its removal reduces the average AUROC from 77.8% to 67.8% and decreases the proportions above 80% and 85% by 33.6 and 22.0 percentage points. Overall, the results support both mechanism-oriented representation learning and reliability-aware aggregation, with view weighting contributing most strongly to zero-shot prediction.

4.5 Representation Space Analysis

We further examine the local structure of the PMRD-center representation by clustering embeddings in the original space and visualizing the resulting enrichment patterns with t-SNE.

As shown in Fig. 4, both datasets exhibit locally coherent regions associated with enriched biological or assay themes. In ChEMBL2K, mechanism-related groups form relatively compact neighborhoods, whereas weakly enriched samples are more dispersed. Broad6K shows a similar pattern, with several assay groups forming distinguishable local structures. These observations suggest that PMRD preserves biologically meaningful and discriminative neighborhood structure, although t-SNE provides only qualitative evidence.

5 Conclusion

In this paper, we present a multimodal framework PMRD for mechanism-aware zero-shot drug property prediction. By separating mechanism-consistent factors from modality-specific information, PMRD constructs a consensus response domain and prioritizes signals that are cross-modally consistent, locally stable, and supportive of discriminative retrieval. Results on ChEMBL2K and Broad6K indicate improved predictive performance and more biologically coherent drug neighborhoods, while hard-negative analysis suggests fewer conflicts between structurally dissimilar but response-related compounds. Our future work will extend PMRD to additional profiling modalities and validate the retrieved mechanism relationships using external biological evidence.

AI USE STATEMENT

In this work, we used generative AI tools solely to improve the readability and language of the manuscript. We did not use generative AI tools for research ideation, experimental design, code generation, data analysis, or the generation of scientific claims. All AI-assisted edits were carefully reviewed and verified by the authors. We take full responsibility for the final content of this work, including any text produced with the aid of generative AI.

REFERENCES

Mark-Anthony Bray, Sigrun M. Gustafsdottir, Mohammad H. Rohban, Shantanu Singh, Vebjorn Ljosa, Katherine L. Sokolnicki, Joshua A. Bittker, Nicole E. Bodycombe, Vlado Dancik, Thomas P. Hasaka, et al. A dataset of images and morphological profiles of 30,000 small-molecule treatments using the cell painting assay. *GigaScience*, 6(12):giw014, 2017.

Srinivas Niranj Chandrasekaran et al. Jump cell painting dataset: Morphological impact of 136,000 chemical and genetic perturbations. *bioRxiv*, 2023.

Steven M. Corsello, Rohith T. Nagari, Ryan D. Spangler, Jordan Rossen, Mustafa Kocak, Jordan G. Bryan, Ranad Humeidi, David Peck, Xiaoyun Wu, Andrew A. Tang, et al. Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. *Nature Cancer*, 1(2): 235–248, 2020.

Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. On uni-modal feature learning in supervised multi-modal learning. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 8632–8656, 2023.

Anna Gaulton, Louisa J. Bellis, A. Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, and John P. Overington. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40(D1):D1100–D1107, 2012.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1263–1272, 2017.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. In *International Conference on Learning Representations*, 2020.

Gwanghoon Jang, Sungjoon Park, Sanghoon Lee, Sunkyu Kim, Sejeong Park, and Jaewoo Kang. Predicting mechanism of action of novel compounds using compound structure and transcriptomic signature coembedding. *Bioinformatics*, 37(Supplement_1):i376–i382, 2021.

Gang Liu, Srijit Seal, John Arevalo, Zhenwen Liang, Anne E. Carpenter, Meng Jiang, and Shantanu Singh. Learning molecular representation in a cell. In *International Conference on Learning Representations*, 2025.

Martin Q. Ma, Yao-Hung Hubert Tsai, Paul Pu Liang, Han Zhao, Kun Zhang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Conditional contrastive learning for improving fairness in self-supervised learning. In *International Conference on Learning Representations*, 2022.

Claire McQuin, Allen Goodman, Vasiliy Chernyshev, Lee Kamentsky, Beth A. Cimini, Kyle W. Karhohs, Minh Doan, Liya Ding, Susanne M. Rafelski, Derek Thirstrup, et al. Cellprofiler 3.0: Next-generation image processing for biology. *PLoS Biology*, 16(7):e2005970, 2018.

Nikita Moshkov, Tim Becker, Kevin Yang, Peter Horvath, Vlado Dancik, Bridget K. Wagner, Paul A. Clemons, Shantanu Singh, Anne E. Carpenter, and Juan C. Caicedo. Predicting compound activity from phenotypic profiles and chemical structures. *Nature Communications*, 14(1):1967, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021a.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021b.

Jiahua Rao, Hanjing Lin, Leyu Chen, Jiancong Xie, Shuangjia Zheng, and Yuedong Yang. Multimodal contrastive learning with negative sampling calibration for phenotypic drug discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30752–30762, 2025.

Joshua Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra. Can contrastive learning avoid shortcut solutions? In *Advances in Neural Information Processing Systems*, volume 34, pp. 4974–4986, 2021.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12559–12571, 2020.

Ana Sanchez-Fernandez, Elisabeth Rumetshofer, Sepp Hochreiter, and Günter Klambauer. CLOOME: Contrastive learning unlocks bioimaging databases for queries with chemical structures. *Nature Communications*, 14(1):7339, 2023a. doi: 10.1038/s41467-023-42328-w.

Ana Sanchez-Fernandez, Elisabeth Rumetshofer, Sepp Hochreiter, and Günter Klambauer. Cloome: Contrastive learning unlocks bioimaging databases for queries with chemical structures. *Nature Communications*, 14(1):7339, 2023b.

Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Liò. 3d infomax improves gnns for molecular property prediction. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 20479–20502, 2022.

Aravind Subramanian, Rajiv Narayan, Steven M. Corsello, David D. Peck, Ted E. Natoli, Xiaodong Lu, Joshua Gould, John F. Davis, Andrew A. Tubelli, Jacob K. Asiedu, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452, 2017.

Chenyu Wang, Sharut Gupta, Caroline Uhler, and Tommi S. Jaakkola. Removing biases from molecular representations via information maximization. In *International Conference on Learning Representations*, 2024.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. Moleculenet: A benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018.

Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of Medicinal Chemistry*, 63(16):8749–8760, 2020.

Minghao Xu, Hang Wang, Bingbing Ni, Hongyu Guo, and Jian Tang. Self-supervised graph-level representation learning with local and global structure. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 11548–11558, 2021.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, 2019.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. In *Advances in Neural Information Processing Systems*, 2020.

Yuning You, Tianlong Chen, Yang Shen, and Zhangyang Wang. Graph contrastive learning automated. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 12121–12132, 2021.

Shuangjia Zheng, Jiahua Rao, Jixian Zhang, Lianyu Zhou, Jiancong Xie, Ethan Cohen, Wei Lu, Chengtao Li, and Yuedong Yang. Cross-modal graph contrastive learning with cellular images. *Advanced Science*, 11(32):2404845, 2024a.

Shuangjia Zheng, Jiahua Rao, Jixian Zhang, Lianyu Zhou, Jiancong Xie, Ethan Cohen, Wei Lu, Chengtao Li, and Yuedong Yang. Cross-modal graph contrastive learning with cellular images. *Advanced Science*, 11(32):2404845, 2024b. doi: 10.1002/advs.202404845.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. In *International Conference on Learning Representations*, 2023.